Supplementary Figure 1. Principle Component Analysis (PCA) of normalized ratios of HGMD vs. SNP substitution for loss or gain of ESE hexamers, respectively. Each row corresponds to a single ESE hexamer whereas each column represents loss or gain of the hexamer by a genomic variant. Each box depicts the log ratio for the counts of HGMD/SNP causing loss or gain of a specific hexamer. A positive log ratio (red) corresponds to a hexamer in a certain context, loss or gain (different columns) that is significantly enriched in inherited disease. Alternatively, a blue value represents a hexamer that is polymorphic across human populations. White boxes correspond to non-significant P-values given a false discovery rate (FDR) of 5%. Regions of interest are designated by a vertical line. Each specific hexamer sequence is corresponding to a row is given on the right of each heat map. These data are identical to those presented in Figure 1C.

Supplementary Figure 2. Principle Component Analysis (PCA) of normalized ratios of HGMD vs. SNP substitution for loss or gain of ESS hexamers, respectively. Each row corresponds to a single ESS hexamer whereas each column represents loss or gain of the hexamer by a genomic variant. Each box depicts the log ratio for the counts of HGMD/SNP causing loss or gain of a specific hexamer. A positive log ratio (red) corresponds to a hexamer in a certain context, loss or gain (different columns) that is significantly enriched in inherited disease. Alternatively, a blue value represents a hexamer that is polymorphic across human populations. White boxes correspond to non-significant P-values given a false discovery rate (FDR) of 5%. Regions of interest are designated by a vertical line. Each specific hexamer sequence is corresponding to a row

is given on the right of each heat map. These data are identical to those presented in Figure 1D.

Supplementary Figure 3. Principle Component Analysis (PCA) of normalized ratios of HGMD vs. SNP substitution for loss or gain of ESE hexamers, respectively. Each row corresponds to a single ESE hexamer whereas each column represents loss or gain of the hexamer by a genomic variant based on their location to the nearest splice site. Each box depicts the log ratio for the counts of HGMD/SNP causing loss or gain of a specific hexamer. A positive log ratio (red) corresponds to a hexamer in a certain context, loss or gain (different columns) that is significantly enriched in inherited disease. Alternatively, a blue value represents a hexamer that is polymorphic across human populations. White boxes correspond to non-significant P-values given a false discovery rate (FDR) of 5%. Regions of interest are designated by a vertical line. Each specific hexamer sequence is corresponding to a row is given on the right of each heat map. Region *i* on the plot corresponds hexamers lost adjacent to 3' splice sites rather than 5' splice sites.

Supplementary Figure 4. Principle Component Analysis (PCA) of normalized ratios of HGMD vs. SNP substitution for loss or gain of ESS hexamers, respectively. Each row corresponds to a single ESS hexamer whereas each column represents loss or gain of the hexamer by a genomic variant based on their location to the nearest splice site. Each box depicts the log ratio for the counts of HGMD/SNP causing loss or gain of a specific hexamer. A positive log ratio (red) corresponds to a hexamer in a certain context, loss or gain (different columns) that is significantly enriched in inherited disease. Alternatively,

a blue value represents a hexamer that is polymorphic across human populations. White boxes correspond to non-significant P-values given a false discovery rate (FDR) of 5%. Regions of interest are designated by a vertical line. Each specific hexamer sequence is corresponding to a row is given on the right of each heat map. Region *i* on the plot corresponds hexamers lost from 3' splice sites. Region *ii* shows hexamers with a similar bias for gain relative to the 3' splice site.

Supplementary Figure 5. Principle Component Analysis (PCA) of normalized ratios of HGMD vs. SNP substitution for loss or gain of ESE hexamers, respectively. Each row corresponds to a single ESE hexamer whereas each column represents loss or gain of the hexamer by a genomic variant based on annotation of the exon (alternative or constitutive). Each box depicts the log ratio for the counts of HGMD/SNP causing loss or gain of a specific hexamer. A positive log ratio (red) corresponds to a hexamer in a certain context, loss or gain (different columns) that is significantly enriched in inherited disease. Alternatively, a blue value represents a hexamer that is polymorphic across human populations. White boxes correspond to non-significant P-values given a false discovery rate (FDR) of 5%. Regions of interest are designated by a vertical line. Each specific hexamer sequence is corresponding to a row is given on the right of each heat map. Region *i* on the plot corresponds hexamers lost from constitutive and non-significant for alternative exons. Region *ii* shows hexamers with a similar bias for gain in constitutive exons.

Supplementary Figure 6. Principle Component Analysis (PCA) of normalized ratios of HGMD vs. SNP substitution for loss or gain of ESS hexamers, respectively. Each row corresponds to a single ESS hexamer whereas each column represents loss or gain of the hexamer by a genomic variant based on annotation of the exon (alternative or constitutive). Each box depicts the log ratio for the counts of HGMD/SNP causing loss or gain of a specific hexamer. A positive log ratio (red) corresponds to a hexamer in a certain context, loss or gain (different columns) that is significantly enriched in inherited disease. Alternatively, a blue value represents a hexamer that is polymorphic across human populations. White boxes correspond to non-significant P-values given a false discovery rate (FDR) of 5%. Regions of interest are designated by a vertical line. Each specific hexamer sequence is corresponding to a row is given on the right of each heat map. Region *i* on the plot corresponds hexamers lost from constitutive and non-significant for alternative exons. Region *ii* shows hexamers with a similar bias for gain in constitutive exons.

Supplementary Figure 7. Distribution of conservation scores for ESE or random hexamers ablated by genomic variants. These data are represented as box plots showing the distribution of average PhyloP scores for distances from 3-36bp from the nearest splice site. The middle boxplots contain distributed average PhyloP scores for 13,000 random hexamers (non-ESRs) sampled from HGMD-targeted exons. The PhyloP distributions for these random hexamers were compared to conservation scores for ESE hexamers ablated by HGMD mutations or SNPs within the region corresponding to 25% of the average exon size nearest to the exon-intron boundary (3-36bp, supplementary

Figure 7). Statistical hypothesis test on means was executed using a Welch t-test using α values of 0.05 (*), 0.01 (**), and 0.001 (***).

Supplementary Figure 8. . Reporter construct sequence and frame context for three splicing reporters. The upstream human β-globin reporter upstream exon sequence with arrows going to each of the individual test exons for *OPA1*, *PYGM* and *TFR2* constructs and the frame context of the exon within the reporter. Each test exon results in the downstream human beta hemoglobin exon being in its endogenous frame. Images extracted from the UCSC Genome Browser. In the case of the optic atrophy 1 (*OPA1*) mutation, a missense (CCA-CTA, Pro-Leu) substitution 14 bp from the 5' splice site serves to create the ACUAGG motif. The mutant exon from the McArdle disease-associated gene muscle glycogen phosphorylase (*PYGM*) isoform 1 contains a nonsense codon (TGG-TAG, Trp-Term) 8 nt upstream of the 5' splice site. Finally, we constructed a reporter based on a nonsense mutation (TAC-TAG, Tyr-Term) in the *TRF2* gene, associated with hemochromatosis type III, which occurs 24bp downstream of the 3' splice.

Supplementary Figure 9. Inhibition of nonsense mediated decay (NMD) by emetine in cells transfected with splicing reporter constructs. HeLa cells were transiently transfected with both wild-type (Wt) and mutant (Mt) alleles from the *TFR2* (panel A), *OPA1* (panel B) and *PYGM* (panel C) constructs. Cells were treated with emetine (+/-) in triplicate. and representative splicing assay of an endogenous NMD-susceptible gene are shown for the *SRSF6* PTC-containing isoform. The PTC-isoform is designated by an

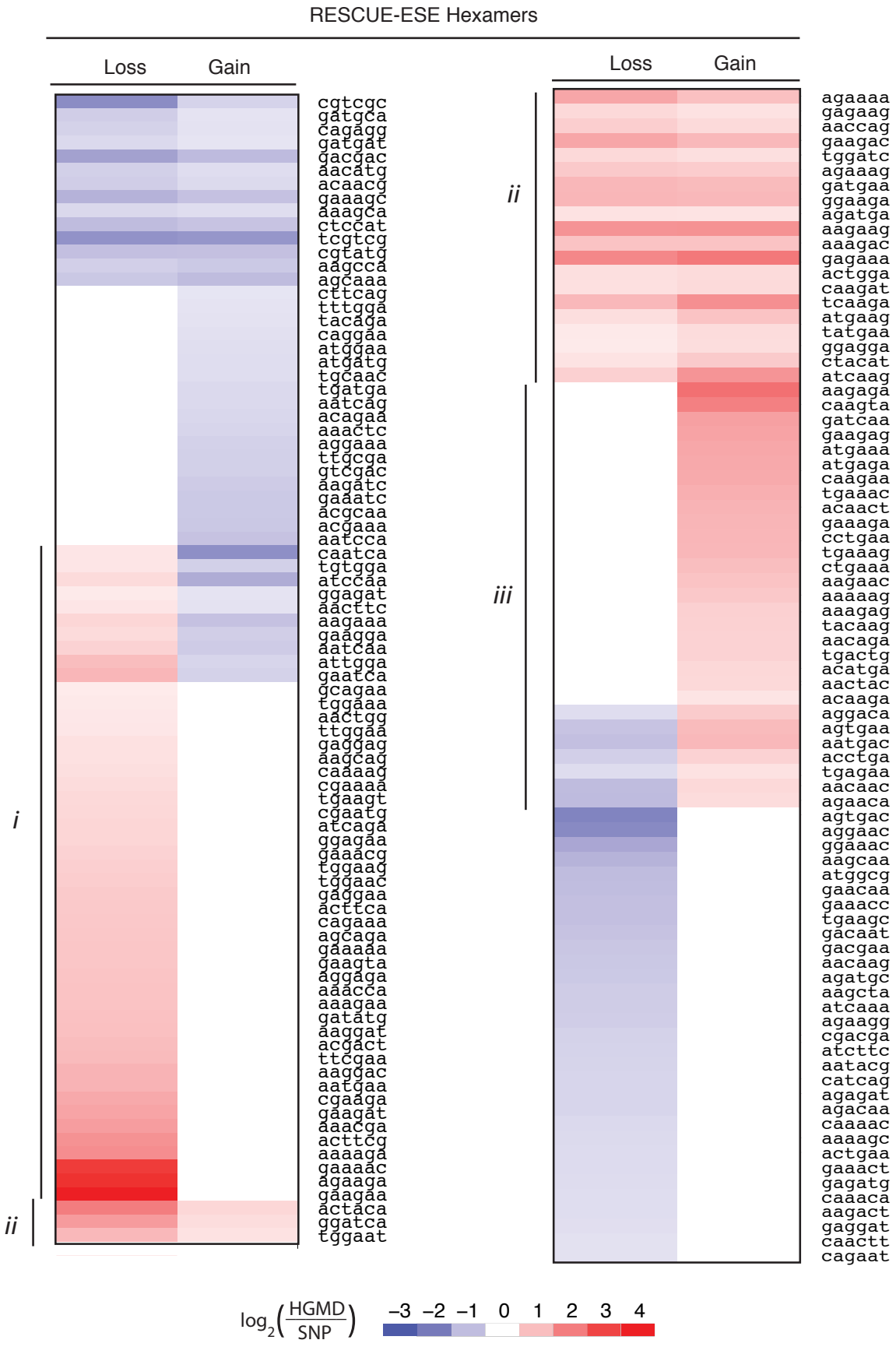
exon with a stop codon. The emetine-dependent PCR product labeled with a star corresponds to a spurious PCR product derived from SRSF6-PTC isoform. cDNA samples corresponding to lanes 7-12 of each panel were used measure the splicing efficiency of reporter in Figure 3.

Supplementary Figure 10. RNA Affinity Chromatography and MudPIT provide an array of candidate *trans*-acting factors. We generated two RNA baits, corresponding to approximately 30 nt of sequence flanking the 5'ss of exon 13 that differ only by a single nucleotide, corresponding to the C>T transition at position -14, relative to the 5'ss. Captured RNA binding proteins from were eluted from the beads by increasing concentrations of KCl, precipitated and analyzed by Multidimensional Protein Identification Technology (MudPIT). (A) Silver stained polyacrylamide gel corresponding to protein elutions at different NaCl concentrations for both the healthy and mutant *OPA1* RNA bait. (B) Specific candidate factors from Supplementary Tables 2 and 3, showing selected results of the MudPIT Mass Spectrometry analysis for 200mM (top) and 1.0M (bottom) KCl elutions.

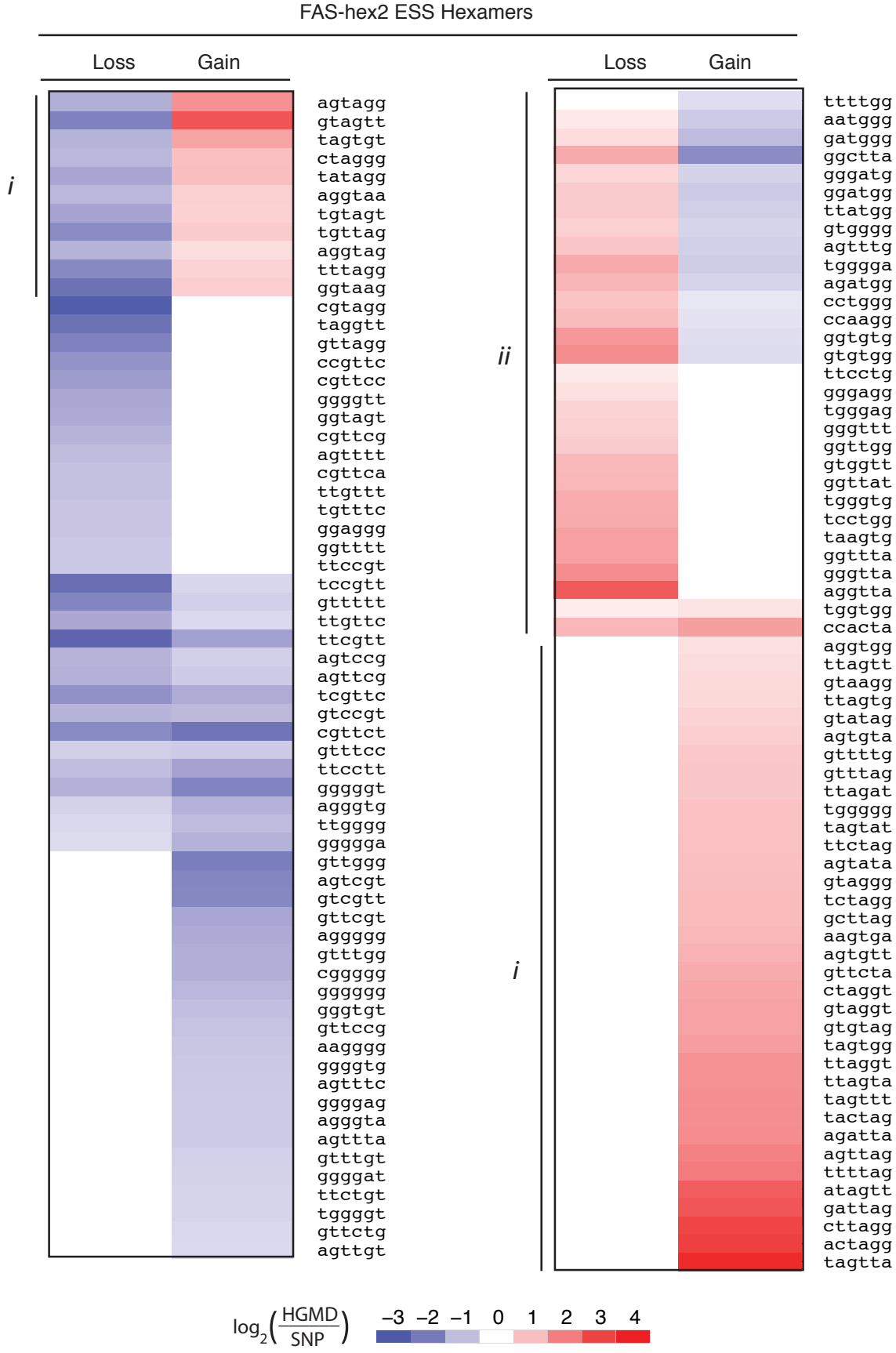
Supplementary Figure 11. Transition/transversion frequencies derived from the HGMD and SNP datasets. (A) Directional allele frequencies for HGMD (with directionality from healthy to disease) mutations from one nucleotide to another. (B) As for panel A, but using the 1000 Genomes Project dataset of SNPs (with directionality from ancestral to variant) for allele frequencies.

Supplementary Figure 12. Titration of PCR cycles for each splicing reporter construct. . HeLa cells were transiently transfected with both wild-type (Wt) and mutant (Mt) alleles from the *TFR2* (panel A), *OPA1* (panel B) and *PYGM* (panel C) constructs. Following treatment with emetine, cytosolic RNA was isolated and converted to cDNA. 66 ng of cDNA was used to template PCR for each splicing reporter. To avoid an PCR bias due to amplification we carefully titrated the number of PCR cycles required to observe reproducible levels of reporter products (29, 32, 35 and 38 cycles). The figure shows a representative gel. For all subsequent experiments from emetine-treated cells 29 cycles of amplification were used to assay splicing efficiencies of the reporters (Figure 3B).

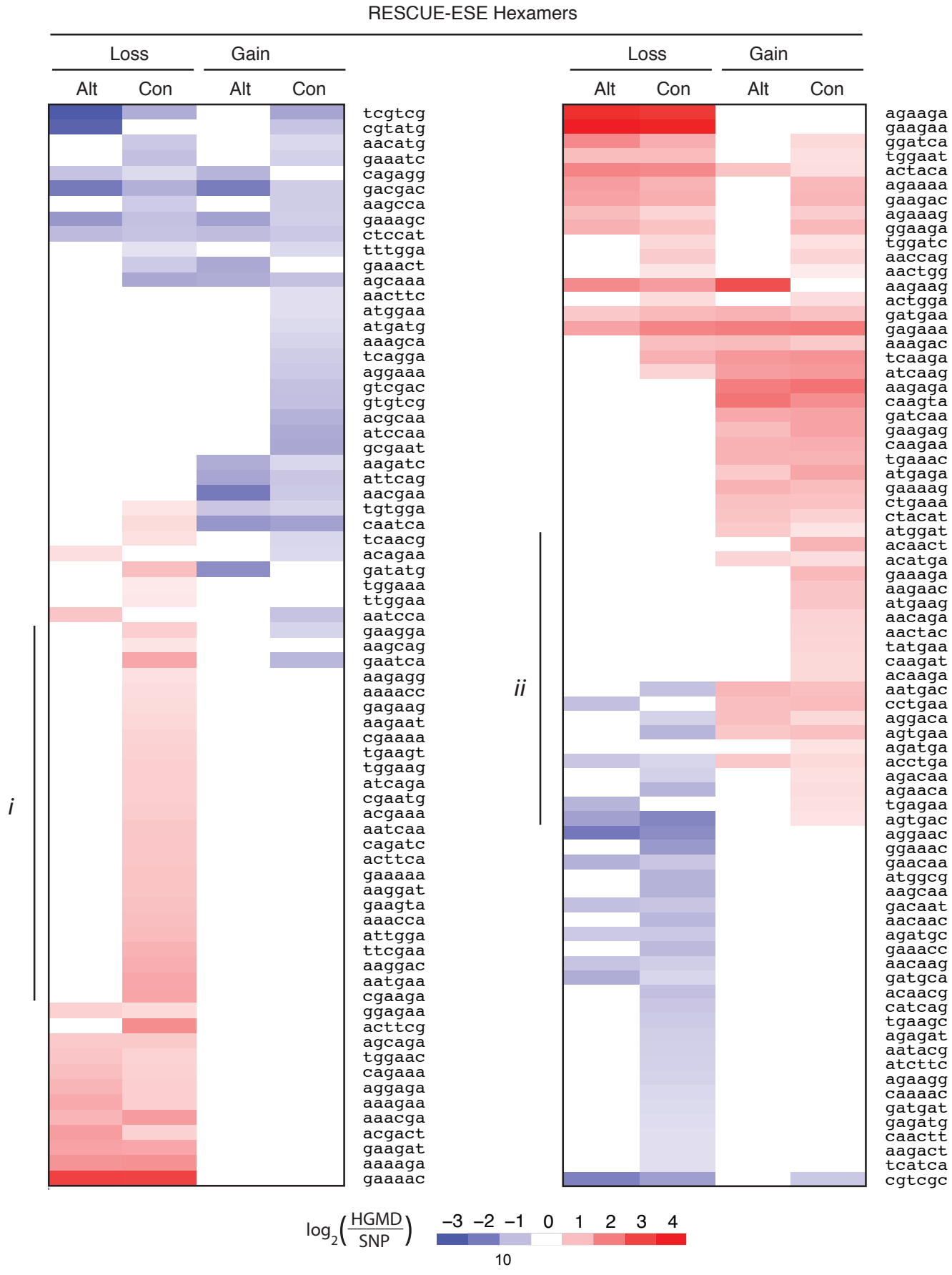
Supplemental Figure 1.



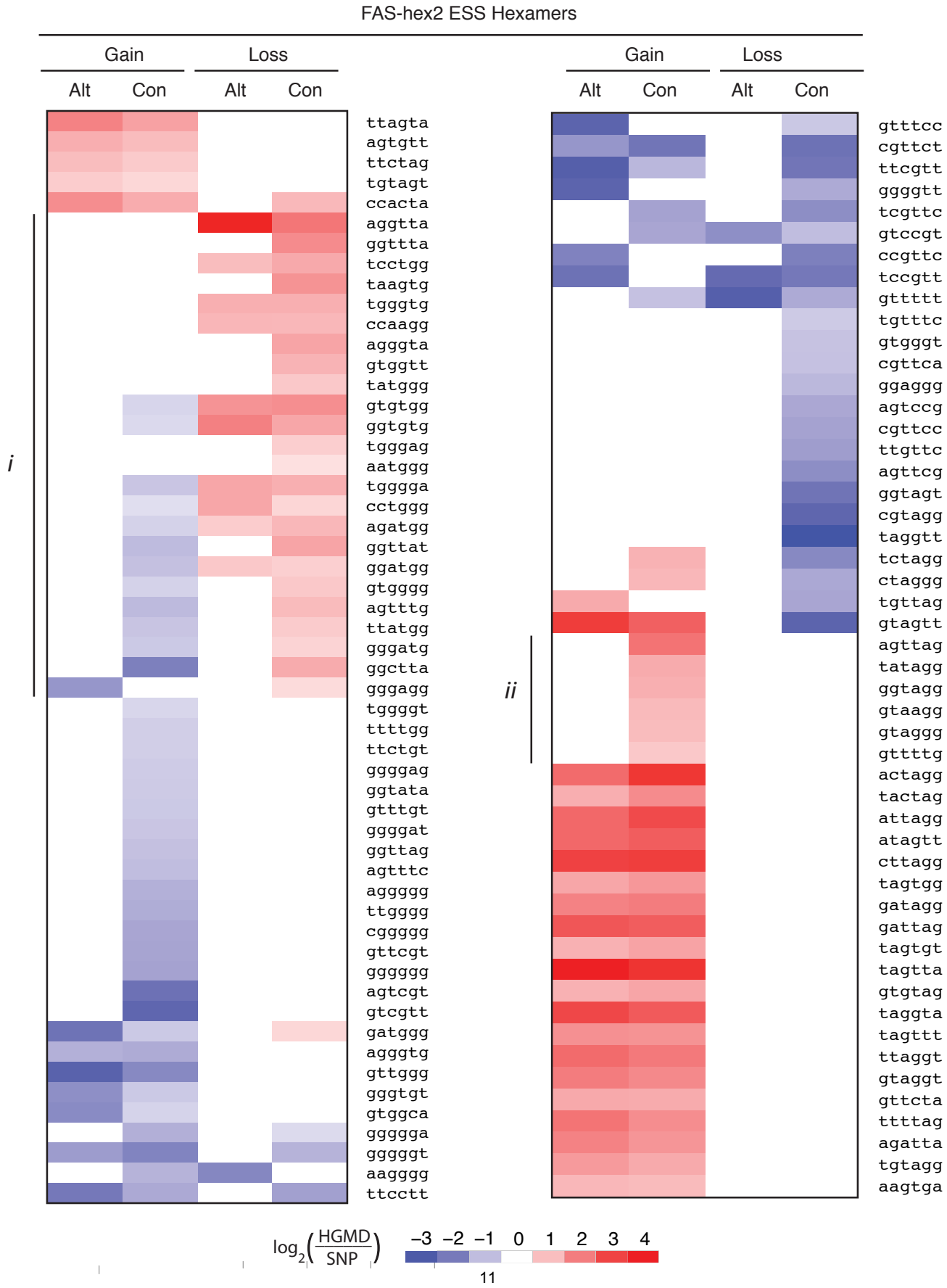
Supplemental Figure 2.



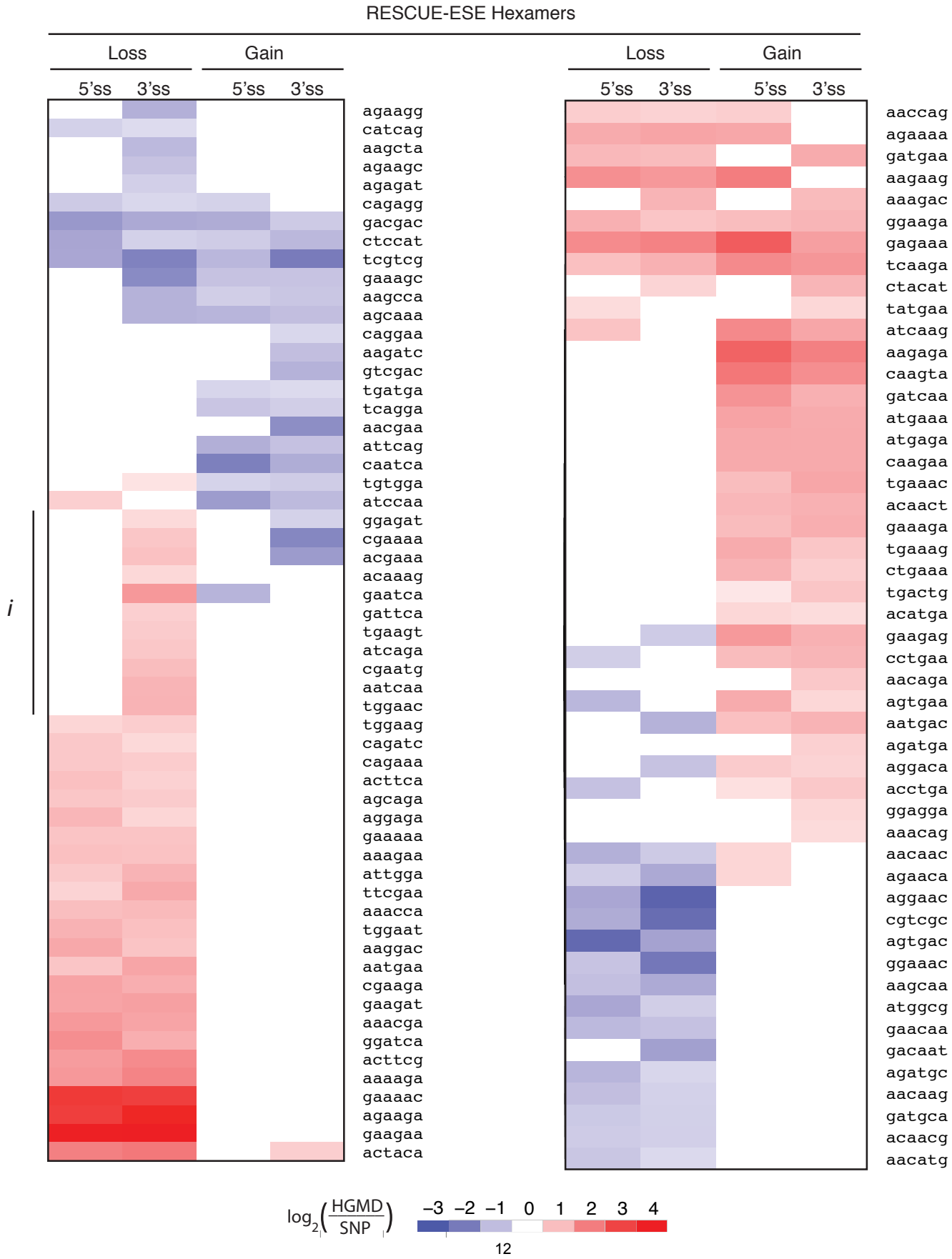
Supplemental Figure 3.



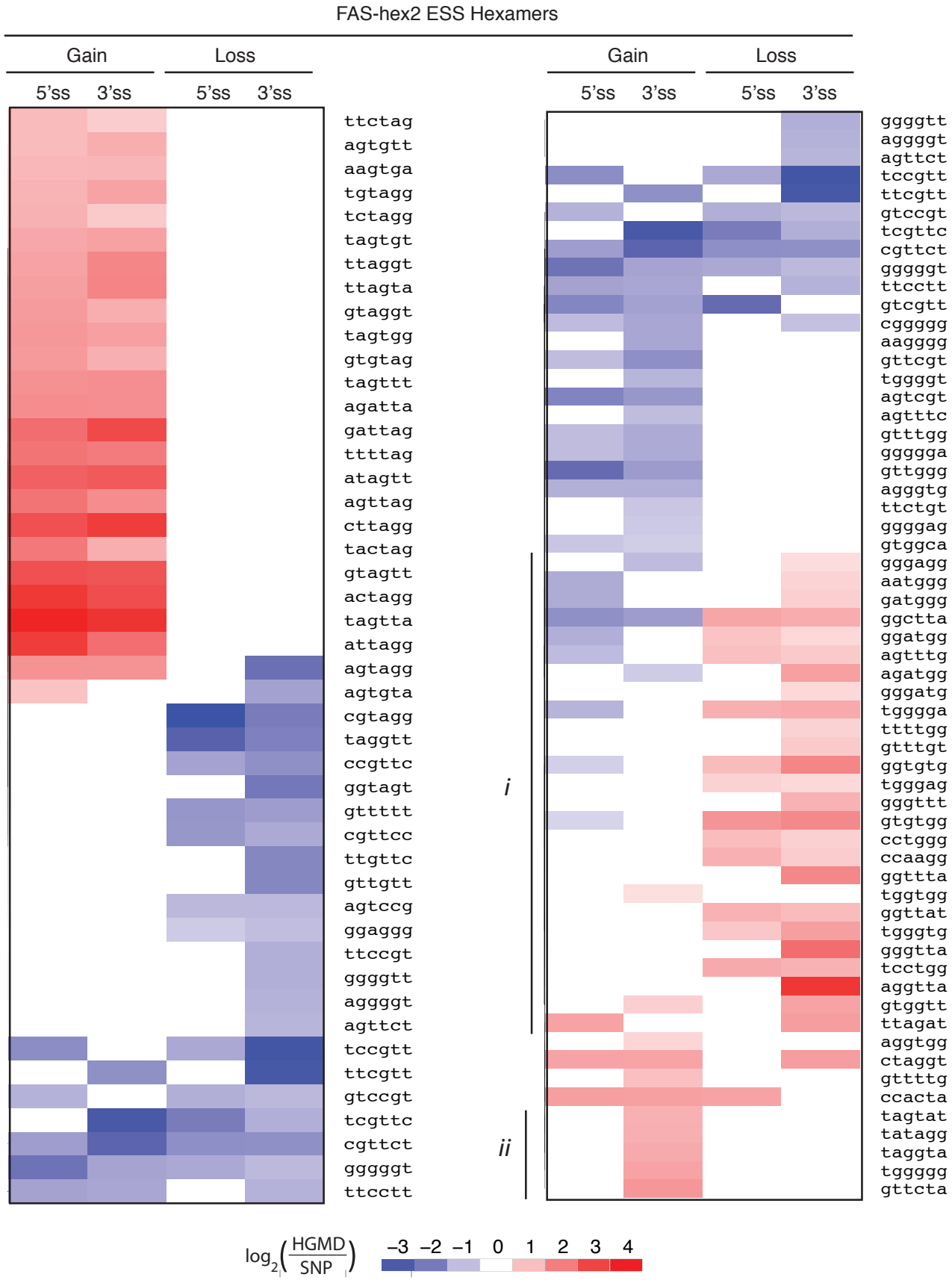
Supplemental Figure 4.



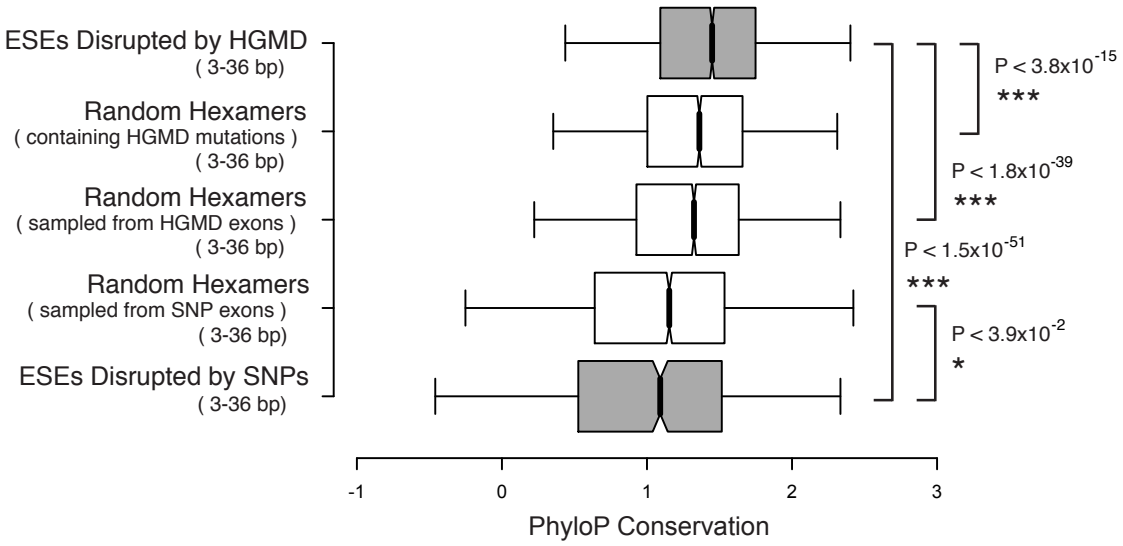
Supplemental Figure 5.



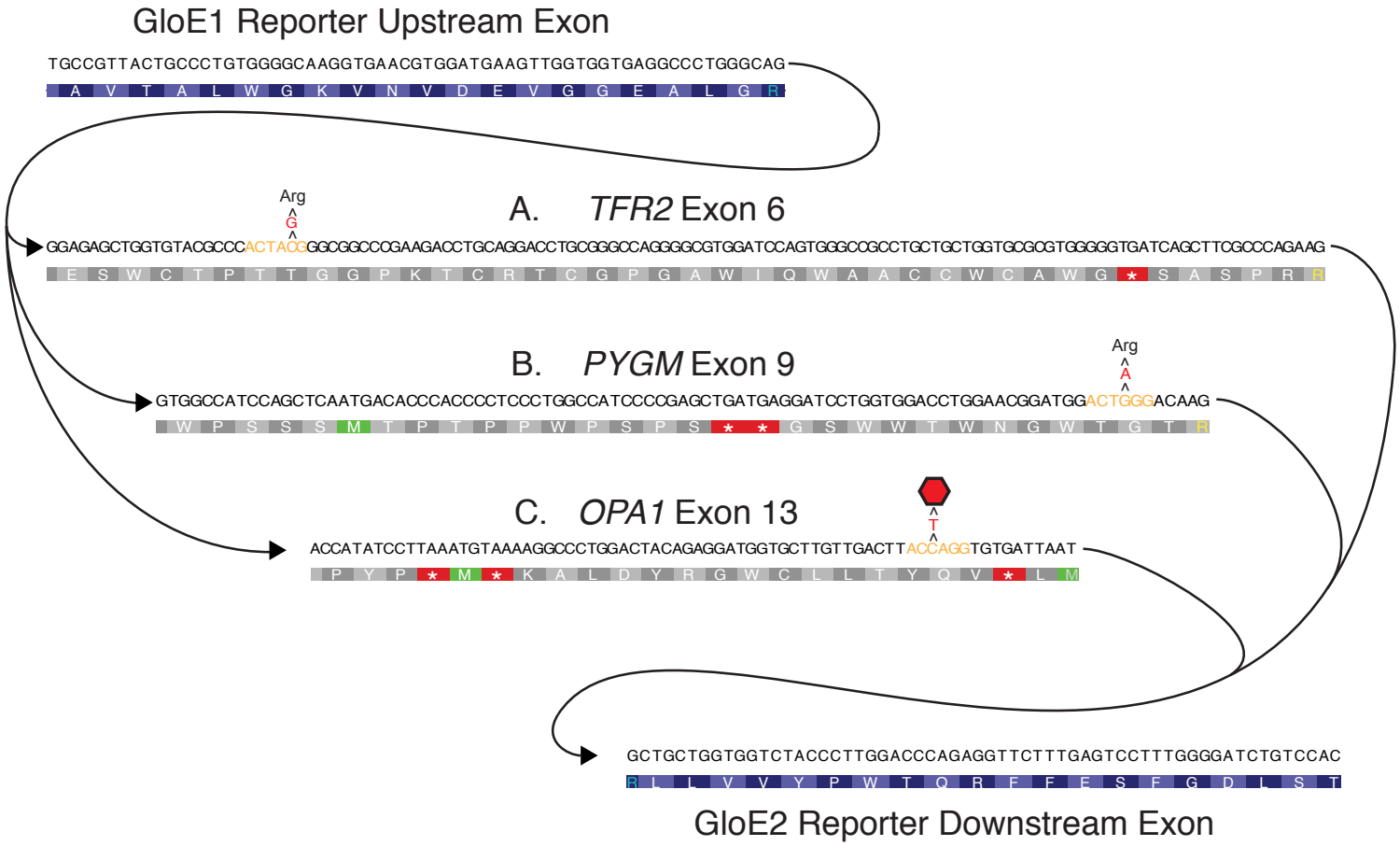
Supplemental Figure 6.



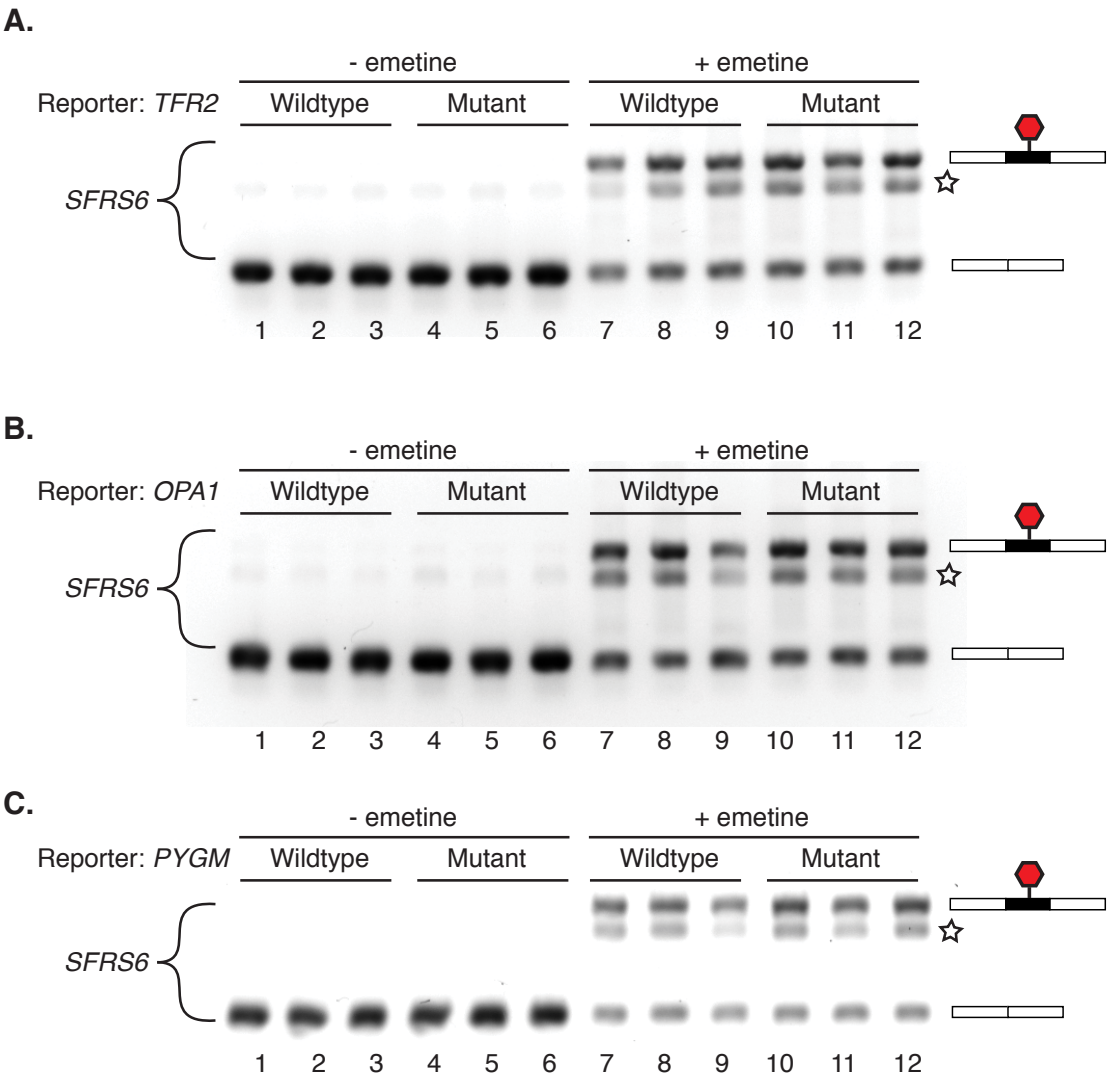
Supplemental Figure 7.



Supplemental Figure 8.

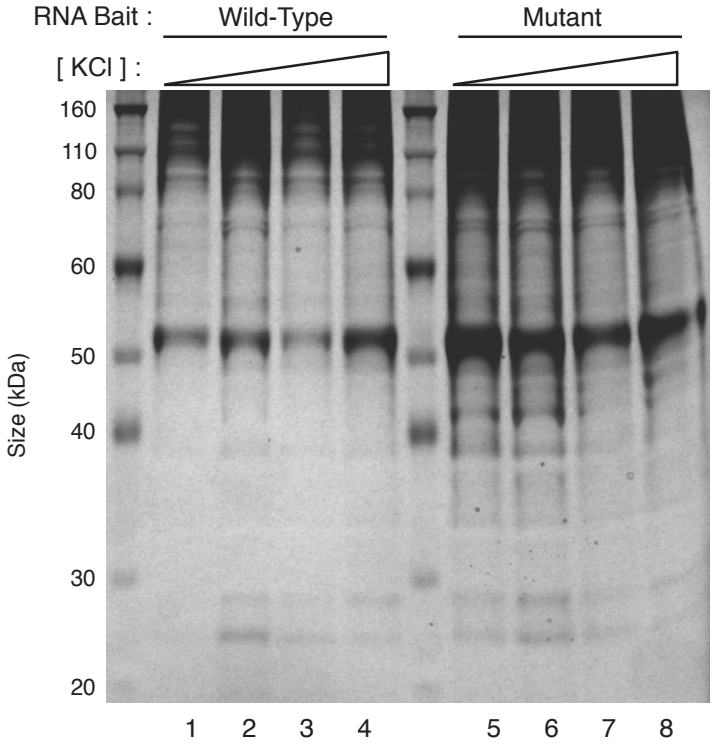


Supplemental Figure 9.

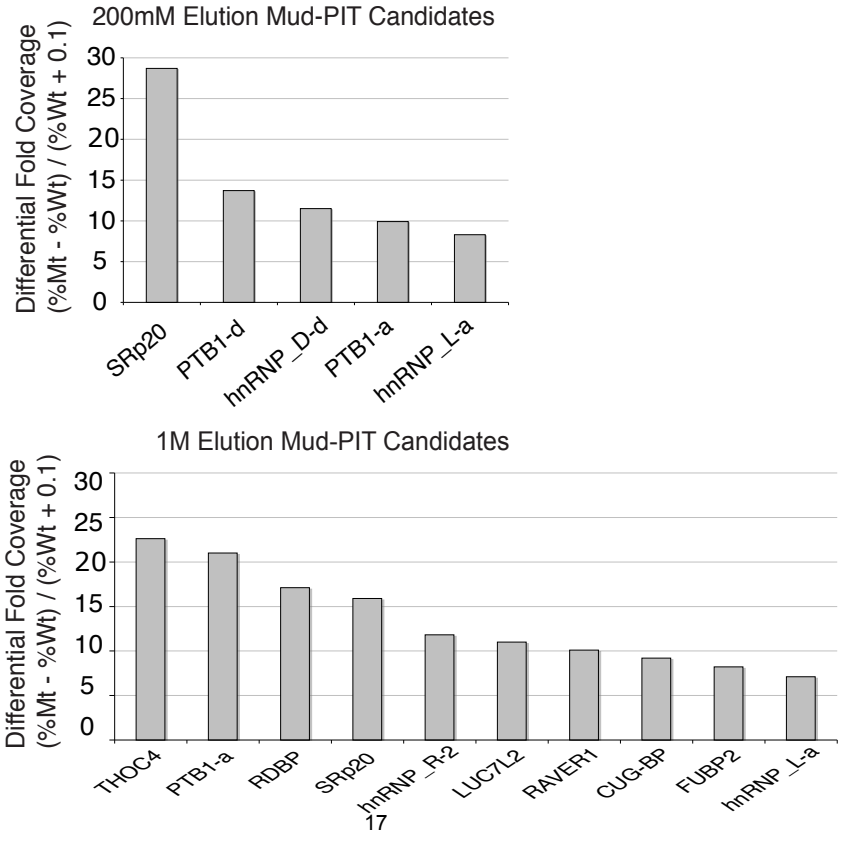


Supplemental Figure 10.

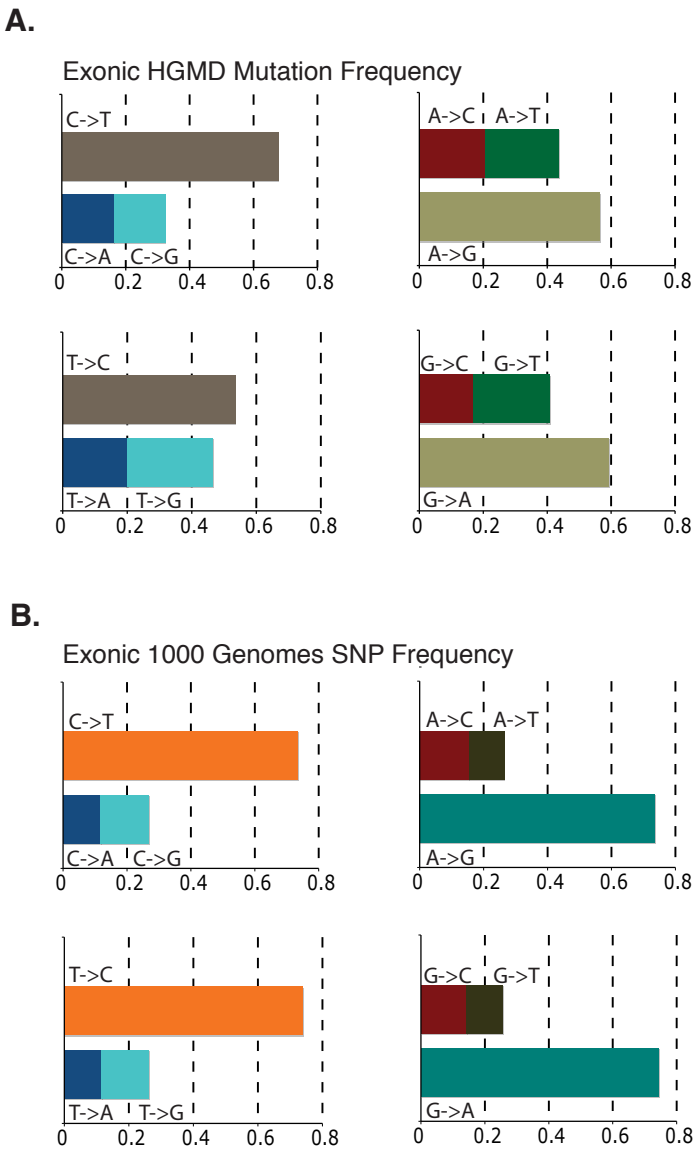
A.



B.



Supplemental Figure 11.



Supplemental Figure 12.